An in-silico framework for comparing and validating transcripts predicted from single and paired-end reads

Anna Paola Carrieri[1,2], Stefano Beretta[1], Gianluca Della Vedova[1], Ernesto Picardi[2], Yuri Pirola[1], Raffaella Rizzi[1] Graziano Pesole[2] and Paola Bonizzoni[1]

[1] Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo),
Università degli studi di Milano-Bicocca, Milano (Italy)
[2] Istituto di Biomembrane e Bioenergetica (IBBE)-CNR, Bari (Italy)

Introduction and motivations

High-throughput sequencing of transcriptome (RNA-Seq) has provided a deep understanding of the complexity of transcriptome and its regulation. Transcriptomes of eukaryotes are characterized by alternative splicing in which multiple isoforms can be produced from a single gene. RNA-Seq analyses estimate that more than 90% of multi-exon human genes are alternatively spliced.

An important issue is characterizing and quantifying alternative splicing events, understanding how these are regulated and how the defects in alternative splicing regulation can influence the onset of human diseases such as spinal muscular atrophy, cystic fibrosis and so on.

RNA-Seq analyses can reveal new genes and new splice variants, as well as quantify expression at genome-wide level. However the volume and the complexity of RNA-Seq data require robust, efficient and sophisticated computational tools to address core data analysis challenges. The still ongoing improvements in RNA-Seq data generation and the research activity towards the analysis of such data has led to rough consensus pipeline consisting of three main steps, that are usually addressed using specific tools. The steps are: read alignment, transcript assembly or genome annotation, transcript and gene expression quantification.

In this context, we have decided to focus our attention on computational methods that use RNA-Seq data to assembly full length mRNA isoforms and estimate their abundance, in order to evaluate and compare their performance and accuracy.

More precisely, in this work we focused on genome-guided assembly methods such as Cufflinks[1] and SLIDE[2]. Both methods adopt an "align-then-assemble" approach, which first relies on computing a spliced alignment of the RNA-Seq reads against a reference genome, then each method reconstructs the transcripts from the spliced alignments. In particular, Cufflinks takes as input a set of aligned reads and assembles the alignments into a parsimonious set of transcripts. Then it estimates their relative abundances based on how many reads support each transcript.

On the other hand, SLIDE (“Sparse Linear modeling of RNA-Seq data for Isoform Discovery and abundance estimation”) uses a statistical method that uses RNA-Seq data to discover mRNA isoforms given an available annotation of gene and exon boundaries, and to estimate the abundance of the predicted isoforms.

Goals

The main aim of our work is to provide a comparison of the previously mentioned methods in order to highlight their strengths and weaknesses in transcript assembly.

For this purpose we designed a software to compare sets of transcripts, each predicted by a different pipeline (the pipelines we have developed are based on Cufflinks or on SLIDE).

The tool is also provided with a visualization feature of the assembled transcripts represented as splicing graphs[3].

Methods and results

We have developed a program for the automatic execution of several pipelines combining different tools dedicated to the main steps of transcript analysis. More specifically, for splice sites alignment and detection, we considered TopHat[4] (a fast splice junction mapper for RNA-Seq reads) and GSNAP[5] (Genomic Short-read Neuclotide Aligment Program). Moreover, to assemble the transcripts, the data generated by the two previous tools are both processed by Cufflinks or SLIDE, providing all possible pipelines.

We have run experiments over two main data sets. Both data are obtained from the annotated isoforms for the set of 112 genes (extracted from the 13 ENCODE regions) used as training set in the EGASP competition[6]. The first experiment consists of the set of single-end reads (1x75bp long) made of all error-free substrings of the annotated isoforms. The second experiment consists of paired-end reads (2x50bp long) obtained with ART[7], which is a simulation tool that generates synthetic next-generation sequencing reads.

In order to have a better understanding of the predictions at different level of detail of input data, we performed two experiments per pipeline per dataset. In the first experiment the reads originating from each gene have been elaborated separately and independently. Instead in the second experiment, all reads have been elaborated together and without any indication of the gene they were originating from. The evaluation of the performance and the accuracy of the two prediction methods has been done using EVAL[8] to compute specificity and sensitivity at transcript level. Furthermore, a more detailed analysis of predicted transcripts at exon level is possible by our visualization tool that synthesize data into splicing graphs. Several results are presented. Among them we highlight that the pipelines combining TopHat with both Cufflinks and SLIDE achieve better specificity and sensitivity. Furthermore, Cufflinks exhibits a greater sensitivity than SLIDE, while SLIDE has a higher specificity, due to the additional input data required by SLIDE w.r.t. Cufflinks.

References

1. C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28(5), 516–520 (May 2010).

2. J. J. Li, C. Jiangb, J. B. Browna, H. Huanga, P. J. Bickel: Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proceedings of the National Academy of Sciences 108(50), 19867-19872 (2011).

3. S. Beretta, P. Bonizzoni, G. Della Vedova, R. Rizzi: Reconstructing Isoform Graphs from RNA-Seq data. IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012), *to appear.*

4. C. Trapnell, L. Pachter, SL. Salzberg: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9), 1105-1111 (2009).

5. T. D. Wu, S. Nacu: Fast and SNP-tolerant detection of complex variants and splicing in short reads: Bioinformatics 26, 873-881 (2010).

6. R. Guigò, P. Flicek, J. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V.B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T.R. Gingeras, J. Harrow, T. Hubbard, S.E. Lewis, M.G. Reese: EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biology 7(1), S2.1–31 (2006).

7. W. Huang, L. Li, J. R. Myers, G. T. Marth: ART: a next-generation sequencing read simulator. Bioinformatics 28(4), 593-594 (2012).

8. E. Keibler, M. R. Brent: Eval: A software package for analysis of genome annotations. BMC Bioinformatics 4(1), 50 (2003).